The emerging science of content labeling: "Soft" interventions and hard public problems

By John P. Wihbey

We are all content labelers — and potentially, the labeled — now. Indeed, we might think of 2020 as the dawn of information about information, the moral use of metadata in the market of speech.

Given this extraordinary turn toward labeling on social media, I want to focus here on a particular research agenda that explores a set of interrelated questions. They revolve around the tricky but fascinating problem of how to label information that may be problematic — incomplete, false, misleading, disputed, or otherwise in need of context. Answering these questions may be key to the organized, efficacious, and ethically justifiable governance of user-generated content on technology platforms, now and long into the future.

There are the narrower, tactical questions that everyone is asking right now about all of the content labeling we have just seen by Twitter, Facebook, and the like. Did any of those 2020 election labels "work"? How about the ones related to COVID-19? How might we define efficacy? How can we improve the user interface and user experience in this regard by tweaking, for example, the features, colors, and interaction design choices of the content labels?

Yet I want to reflect on deeper questions that are beginning to emerge around two areas: ethics and epistemology. These have come into focus as content labeling efforts have continually seemed haphazard, reactive, and often contradictory.

First, there are ethical quandaries that are only beginning to be addressed: How can content labeling become something less tactical and more strategic, systematically linked to thoughtful principles? How can it be grounded in strong ethical norms about how to treat users? On what ethical grounds can social media companies proceed? How can user groups and third-party entities such as news organizations lend independence, legitimacy, and authority to these efforts?

Beyond this, there are core questions about knowledge. How can content labeling efforts improve the epistemic position of platform users, i.e., their ability to form good beliefs about the quality of the information with which they interact on the platform? How can these efforts appropriately respect sources and subjects of information? How can we think about "boosting" users in addition to nudging them?

Accelerating friction

The COVID-19 pandemic accelerated the labeling trend in the early part of the year, as companies used more aggressive fact-checking and moderation techniques, scaled through algorithms. By the end of 2020, everyone from fringe conspiracy theorists to the sitting U.S. president saw their messages on social media being labeled by content moderators as disputed, false, fact-checked, or otherwise in need of further contextual and truthful information. Facebook reportedly labeled 180 million pieces of content during the election season.

The practice of content labeling, of course, was greeted by intellectually contorted howls of "censorship" by the likes of President Donald Trump and his allies. Others saw the measures as a sensible compromise between the past attitude of technolibertarian laissez-faire on the one hand, and draconian takedowns and Orwellian thought-policing. And still another group thought the whole effort was inadequate to slay the misinformation and disinformation dragon.

And no one, it seems, knows how effective any of this labeling ultimately is. Twitter, for its part, has stated that its 2020 election-related labels limited the use of certain kinds of shares, "due in part to a prompt that warned people prior to sharing." Critics maintain that algorithmic downranking must accompany these moves, and too many people still see the misinformation before it is labeled.

From the time that the contemporary social media companies first came on the scene (Facebook in 2004, YouTube in 2005, and Twitter in 2006) until very recently, these companies had either not thought much at all about the health, safety, and integrity of their platforms, or they had come to believe that certain kinds of law- or norm-violating speech (e.g., incitements to violence, IP violations, child pornography) required a blunt remedy. This meant either takedowns ("removal"), or algorithmic reduction in visibility, making content scarcely visible in feeds or timelines.

Alongside removal and reduction, a new treatment arose more recently — namely, "information" treatments, or "friction," "context," or general "disclose" conditions. These are considered "soft" treatments, ones less potentially violating of user rights and freedom of expression. It's an evolving vocabulary that very much depends on the particular company or researcher involved. This general impulse is manifested in the use of labels, interstitials, panels, warning signs, and other treatments that often put the equivalent of scare quotes around content. Related tactics such as transparency pages and source information have arisen in parallel. Companies seemed to have been reading Richard Thaler's and Cass Sunstein's "Nudge" and basically operationalizing it

for the age of social media dilemmas, trying to improve decisions of users and slow the virality of certain kinds of falsehoods.

In 2020, a year that was a "Super Bowl of misinformation," companies such as Twitter, Facebook, and TikTok all accelerated these kinds of labeling efforts, while others such as YouTube took a slower approach to the practice. The Election Integrity Partnership has very usefully and precisely documented many of the associated platform policy changes, which were diverse and often, taken together, incoherent.

The Content Labeling Project

The research questions that are opened up are legion; they are incredibly varied in the fields they might draw on (psychology, philosophy, linguistics and semiotics, information design and visualization, sociology, and political science, to name a few). There are now a range of vital questions to which researchers, journalists, platform users, and of course the companies themselves need answers.

Our project at the Northeastern University Ethics Institute is taking an "information ethics" approach to relevant questions, drawing on the deeper resources of philosophy, particularly the field of social epistemology, to help guide content moderation and labeling practices. (Our project is independent of any particular platform effort, although we have support from Facebook, and I advise some of Twitter's efforts.) Social epistemology has grown as a field in recent decades, and the ways that it approaches questions of knowledge and information seem particularly apt and useful in our connected age.

We aim to abstract away from the moderation tactics du jour and focus on enduring, core questions of ethics and epistemology, laying out what we hope is a solid framework through which content moderators on any sociotechnical platform might approach all contextualization and labeling problems. We want to help the field think through an overarching approach, one based on a clear conception of the point of the strategy and the values, or normative considerations, that the strategy is meant to accomplish or that guide it. Our research is taking inspiration and insights from fields such as nutrition labeling and library science, fields that have long thought about labeling questions.

We are also conducting online experiments to try to answer some of the deeper questions about information correction, a literature that has been accumulating, not always in linear fashion, for more than a decade now. The first in a series of working papers and reports from our research project, co-authored by Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, and myself, is just out. It

is a literature review on a variety of questions related to this emerging science of content labeling. There are a bewildering variety of behavioral and cognitive phenomena to be considered.

In the area of psychological effects, would-be labelers and platform policy managers should know about: the illusory truth effect; the “backfire” effect; the continued influence effect; and the implied truth effect. Some of these are more worrisome than others. But we see an important subdomain of literature having developed that is vital for framing intelligent content moderation decisions. There is also important and relevant research literature around more tactical issues, such as aesthetic characteristics, graphics, alternative media formats, levels of detail, and named source considerations.

What all of this points to is the need for more research around shared questions that speak to a new moment in our networked information society. We are quickly moving away from the controlling ideas for news and information of the 20th century, embodied in former Supreme Court Justice Oliver Wendell Holmes’ notion that ultimate goods are produced by the “free trade in ideas” within the “competition of the market.” From the prevailing idea of *competition* in the marketplace of ideas, we are moving to a paradigm where *orientation* in the marketplace of ideas is becoming paramount. Scale, algorithms, and network effects all are pushing us in that direction.

Content labeling is one logical place for research to focus. The push for greater orientation is an intellectual undertaking that will take large-scale experimentation, as we saw in 2020, as well as much careful and critical thinking in the years ahead.

John P. Wihbey is a faculty affiliate with the Ethics Institute at Northeastern University, an assistant professor of journalism and media innovation, and author of “The Social Fact: News and Knowledge in a Networked World.”